PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12N 15/10, 15/31, C07K 14/395, C12Q 1/68, 1/02 — A2

(11) International Publication Number: **WO 98/32847**

(43) International Publication Date: 30 July 1998 (30.07.98)

(21) International Application Number: PCT/US98/01216

(22) International Filing Date: 22 January 1998 (22.01.98)

(30) Priority Data:
60/035,917 23 January 1997 (23.01.97) US

(71) Applicant: THE JOHNS HOPKINS UNIVERSITY SCHOOL OF MEDICINE [US/US]; 720 Rutland Avenue, Baltimore, MD 21205 (US).

(72) Inventors: VELCULESCU, Victor, E.; Apartment C, 650 North Calvert Street, Baltimore, MD 21202 (US). VOGELSTEIN, Bert; 3700 Breton Way, Baltimore, MD 21208 (US). KINZLER, Kenneth, W.; 1348 Halstead Road, Baltimore, MD 21234 (US).

(74) Agents: KAGAN, Sarah, A. et al.; Banner & Witcoff, Ltd., 11th floor, 1001 G Street, N.W., Washington, DC 20001-4597 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, GW, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: CHARACTERIZATION OF THE YEAST TRANSCRIPTOME

(57) Abstract

Yeast genes which are differentially expressed during the cell cycle are described. They can be used to study, affect, and monitor the cell cycle of a eukaryotic cell. They can be used to obtain human homologs involved in cell cycle regulation. They can be used to identify antifungal agents. They can be formed into arrays on solid supports for interrogation of a cell's transcriptome under various conditions.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

CHARACTERIZATION OF THE YEAST TRANSCRIPTOME

TECHNICAL FIELD OF THE INVENTION

This invention is related to the characterization of the expressed genes of the yeast genome. More particularly, it is related to the identification and use of previously unrecognized genes.

BACKGROUND OF THE INVENTION

It is by now axiomatic that the phenotype of an organism is largely determined by the genes expressed within it. These expressed genes can be represented by a "transcriptome", conveying the identity of each expressed gene and its level of expression for a defined population of cells. Unlike the genome, which is essentially a static entity, the transcriptome can be modulated by both external and internal factors. The transcriptome thereby serves as a dynamic link between an organism's genome and its physical characteristics.

The transcriptome as defined above has not been characterized in any eukaryotic or prokaryotic organism, largely because of technological limitations. However, some general features of gene expression patterns were elucidated two decades ago through RNA-DNA hybridization measurements (Bishop et al., 1974; Hereford and Rosbash, 1977). In many organisms, it was thus found that at least three classes of transcripts could be identified, with either high, medium, or low levels of expression, and the

number of transcripts per cell were estimated (Lewin, 1980). These data of course provided little information about the specific genes that were members of each class. Data on the expression levels of individual genes have accumulated as new genes were discovered. However, in only a few instances have the absolute levels of expression of particular genes been measured and compared to other genes in the same cell type.

Description of any cell's transcriptome would therefore provide new information useful for understanding numerous aspects of cell biology and biochemistry.

SUMMARY OF THE INVENTION

It is an object of the present invention to provide genes which are involved in cell cycle progression.

It is another object of the present invention to provide methods of using the genes to affect the cell cycle.

It is an object of the present invention to provide methods for screening candidate antifungal drugs.

Another object of the invention is to provide a method for obtaining human homologs of the yeast genes which are involved in cell cycle progression.

Another object of the invention is to provide probes for ascertaining phase in the cell cycle of a cell.

These and other objects of the invention are achieved by providing the art with one or more of the embodiments described below. According to one embodiment of the invention an isolated DNA molecule is provided. It comprises a yeast gene which is involved in cell cycle progression selected from the group of NORF genes identified in Table 3 or 4.

According to another embodiment of the invention a method of using yeast genes is provided. The method is for affecting the cell cycle of a cell. The method comprises the step of:

administering to a cell an isolated DNA molecule comprising a

yeast gene which is involved in cell cycle progression selected from the differentially expressed genes identified in Tables 1, 2, 3 and 4.

In yet another embodiment of the invention a method for screening candidate antifungal drugs is provided. The method comprises the steps of:

contacting a test substance with a yeast cell;

monitoring expression of a yeast gene which is involved in cell cycle progression selected from the group of yeast genes identified in Tables 1, 2, 3 and 4, wherein a test substance which modifies the expression of the yeast gene is a candidate antifungal drug.

In still another embodiment of the invention a method for identifying human genes which are involved in cell cycle progression is provided. The method comprises the step of:

hybridizing a probe comprising at least 14 contiguous nucleotides of a yeast gene which is differentially expressed between at least two phases selected from the group consisting of log phase, S phase, and G2/M phase, wherein the yeast gene is identified in Table 1, 2, 3, or 4.

Also provided by the present invention are isolated DNA molecules, which comprise probes for ascertaining phase in the cell cycle of a cell, wherein the probe comprises at least 14 contiguous nucleotides of a NORF gene as identified in Table 3 or 4.

These and other embodiments of the invention which will be apparent to those of skill in the art upon reading the detailed disclosure provided below, make available to the art hitherto unrecognized genes, and information about the expression of genes globally at the organismal level. We provide the first description of a transcriptome, determined in *S. cerevisiae* cells. This organism was chosen because it is widely used to clarify the biochemical and physiologic parameters underlying eukaryotic cellular functions and because it is the only eukaryote in which the entire genome has been defined at the nucleotide level (Goffeau, et al., 1996).

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1. Schematic of SAGE Method and Genome Analysis.

In applying SAGE to the analysis of yeast gene expression patterns, the 3' most NlaIII site was used to define a unique position in each transcript and to provide a site for ligation of a linker with a BsmFI site. The type IIs enzyme BsmFI, which cleaves a defined distance from its non-palindromic recognition site, was then used to generate a 15bp SAGE tag (designated by the black arrows), which includes the NlaIII site. Automated sequencing of concatenated SAGE tags allowed the routine identification of about a thousand tags per sequencing gel. Once sequenced, the abundance of each SAGE tag was calculated, and each tag was used to search the entire yeast genome to identify its corresponding gene. The lower panel shows a small region of Chromosome 15. Gray arrows indicate all potential SAGE tags (NlaIII sites) and black arrows indicate 3' most SAGE tags. The total number of tags observed for each potential tag is indicated above (+ strand) or below (- strand) the tag. As expected, the observed SAGE tags were associated with the 3' end of expressed genes.

Figure 2. Sampling of Yeast Gene Expression.

Analysis of increasing amounts of ascertained tags reveals a plateau in the number of unique expressed genes. Triangles represent genes with known functions, squares represent genes predicted on the basis of sequence information, and circles represent total genes.

Figure 3. Virtual Rot.

(a) Abundance Classes in the Yeast Transcriptome. The transcript abundance is plotted in reverse order on the abscissa, whereas the fraction of total transcripts with at least that abundance is plotted on the ordinate. The dotted lines identify the three components of the curve, 1, 2, and 3. This is analogous to a Rot curve derived from reassociation kinetics where the product of initial RNA concentration and time is plotted on the abscissa, and

the percent of labeled cDNA that hybridizes to excess mRNA is plotted on the ordinate.

(b) Comparison of Virtual Rot and Rot Components. Transitions and data from virtual Rot components were calculated from the data in Figure 3A, while data for Rot components were obtained from Hereford and Rosbash, 1977.

Figure 4. Chromosomal Expression Map for S. cerevisiae. Individual yeast genes were positioned on each chromosome according to their open reading frame (ORF) start coordinates. Abundance levels of tags corresponding to each gene are displayed on the vertical axis, with transcription from the + strand indicated above the abscissa and that from the - strand indicated below. Yellow bands at ends of the expanded chromosome represent telomeric regions that are undertranscribed (see text for details).

Figure 5. Northern Blot Analysis of Representative Genes. TDH2/3, TEF1/2 and NORF1, are expressed relatively equally in all three states (lane 1, G2/M arrested; lane 2, S phase arrested; lane 3, log phase), while RNR4, RNR2 , and NORF5 are highly expressed in S-phase arrested cells. The expression level observed by SAGE (number of tags) is noted below each lane and was highly correlated with quantitation of the Northern blot by PhosphorImager analysis (r^2=0.97).

Table Legends

Table 1. Highly Expressed Genes

Tag represents the 10 bp SAGE tag adjacent to the NlaIII site; Gene represents the gene or genes corresponding to a particular tag (multiple genes that match unique tags are from related families, with an average identity of 93%); Locus and Description denote the locus name, and functional description of each ORF, respectively; Copies/cell represents the abundance of each transcript in the SAGE library, assuming 15,000 total transcripts per cell and 60,633 ascertained transcripts.

Table 2. Expression of Putative Coding Sequences

Table columns are the same as for Table 1.

Table 3. Expression of NORF genes

SAGE Tag, Locus, and Copies/cell are the same as for Table 1; Chr and Tag Pos denote the chromosome and position of each tag; ORF Size denotes the size of the ORF corresponding to the indicated tag. In each case, the tag was located within or less than 250 bp 3' of the NORF.

DETAILED DESCRIPTION

It is a discovery of the present invention that certain hitherto unknown genes (the NORFs) exist and are expressed in yeast. These genes, as well as other previously identified and previously postulated genes, can be used to study, monitor, and affect phase of cell cycle. The present invention provides information on which genes are differentially expressed during the cell cycle. Differentially expressed genes can be used as markers of phases of the cell cycle. They can also be used to affect a change in the phase of the cell cycle. In addition, they can be used to screen for drugs which affect the cell cycle, by affecting expression of the genes. Human homologs of these eukaryotic genes are also presumed to exist, and can be identified using the yeast genes as probes or primers to identify the human homologs.

New genes termed NORFs (not previously assigned open reading frames) have been found. They are uniquely identified by their SAGE tags. In addition their entire nucleotide sequence is known and publicly available. In general, these were not previously identified as genes due to their small size. However, they have now been found to be expressed.

Differentially expressed yeast genes are those whose expression varies by a statistically significant difference (to greater than 95% confidence level) within different growth phases, particularly log phase, S phase, and G2/M. Preferably the difference is greater than 10%, 25%, 50%, or 100%. The genes which have been found to have such differential expression characteristics are: NORF Nº 1, 2, 4, 5, 6, 17, 25, 27, TEF1/TEF2, EN02, ADH1, ADH2, PGK1, CUP1A/CUP1B, PYK1, YKL056C, YMR116C, YEL033W, YOR182C, YCR013C, ribonucleotide reductase 2 and 4, and YJR085C.

The DNA molecules according to the invention can be genomic or cDNA. Preferably they are isolated free of other cellular components such as membrane components, proteins, and lipids. They can be made by a cell and isolated, or synthesized using PCR or an automatic synthesizer. Any technique for obtaining a DNA of known sequence may be used. Methods for purifying and isolating DNA are routine and are known in the art.

To administer yeast genes to cells, any DNA delivery techniques known in the art may be used, without limitation. These include liposomes, transfection, transduction, transformation, viral infection, electroporation. Vectors for particular purposes and characteristics can be selected by the skilled artisan for their known properties. Cells which can be used as gene recipients are yeast and other fungi, mammalian cells, including humans, and bacterial cells.

Antifungal drugs can be identified using yeast cells as described herein. Expression of a differentially expressed gene can be monitored by any means known in the art. When a test substance affects the expression of such a differentially expressed gene, it is a candidate drug for affecting the growth

properties of fungi, and may be useful as an antifungal agent.

Because differentially expressed genes are likely to be involved in cell cycle progression, it is likely that these genes are conserved among species. The differentially expressed genes identified by the present invention can be used to identify homologs in humans and other mammals. Means for identifying homologous genes among different species are well known in the art. Briefly, stringency of hybridization can be reduced so that imperfectly matching sequences hybridize. This can be in the context of *inter alia* Southern blots, Northern blots, colony hybridization or PCR. Any hybridization technique which is known in the art can be used.

Probes according to the present invention are isolated DNA molecules which have at least 10, and preferably at least 12, 14, 16, 18, 20, or 25 contiguous nucleotides of a particular NORF gene or other differentially expressed gene. The probes may or may not be labeled. They may be used as primers for PCR or for Southern or Northern blots. Preferably the probes are anchored to a solid support. More preferably they are present on an array so that multiple probes can simultaneously hybridize to a single biological sample. The probes can be spotted onto the array or synthesized *in situ* on the array. See Lockhart et. al., **Nature Biotechnology**, Vol. 14, December 1996, "Expression monitoring by hybridization to high-density oligonucleotide arrays." A single array can contain more than 100, 500 or even 1,000 different probes in discrete locations.

The above disclosure generally describes the present invention. A more complete understanding can be obtained by reference to the following specific examples which are provided herein for purposes of illustration only, and are not intended to limit the scope of the invention.

EXAMPLE

Summary

We have analyzed the set of genes expressed from the yeast genome, herein

called the transcriptome, using serial analysis of gene expression (SAGE). Analysis of 60,633 transcripts revealed 4,665 genes, with expression levels ranging from 0.3 to over 200 transcripts per cell. Of these genes, 1,981 had known functions, while 2,684 were previously uncharacterized. Integration of positional information with gene expression data allowed the generation of chromosomal expression maps, identifying physical regions of transcriptional activity, and identified genes that had not been predicted by sequence information alone. These studies provide insight into global patterns of gene expression in yeast and demonstrate the feasibility of genome-wide expression studies in eukaryotes.

Results

Characteristics and Rationale of SAGE Approach

Several methods have recently been described for the high throughput evaluation of gene expression (Nguyen et al., 1995; Schena et al., 1995; Velculescu et al., 1995). We used SAGE (Serial Analysis of Gene Expression) because it can provide quantitative gene expression data without the prerequisite of a hybridization probe for each transcript. The SAGE technology is based on two basic principles (Figure 1). First, a short sequence tag (9-11 bp) contains sufficient information to uniquely identify a transcript, provided that it is derived from a defined location within that transcript. Second, many transcript tags can be concatenated into a single molecule and then sequenced, revealing the identity of multiple tags simultaneously. The expression pattern of any population of transcripts can be quantitatively evaluated by determining the abundance of individual tags and identifying the gene corresponding to each tag.

Genome-wide expression

In order to maximize representation of genes involved in normal growth and cell-cycle progression, SAGE libraries were generated from yeast cells in three states: log phase, S phase arrested and G2/M phase arrested. In total,

SAGE tags corresponding to 60,633 total transcripts were identified (including 20,184 from log phase, 20,034 from S phase arrested, and 20,415 from G2/M phase arrested cells). Of these tags, 56,291 tags (93%) precisely matched the yeast genome, 88 tags matched the mitochondrial genome, and 91 tags matched the 2 micron plasmid.

The number of SAGE tags required to define a yeast transcriptome depends on the confidence level desired for detecting low abundance mRNA molecules. Assuming the previously derived estimate of 15,000 mRNA molecules per cell (Hereford and Rosbash, 1977), 20,000 tags would represent a 1.3 fold coverage even for mRNA molecules present at a single copy per cell, and would provide a 72% probability of detecting such transcripts (as determined by Monte Carlo simulations). Analysis of 20,184 tags from log phase cells identified 3,298 unique genes. As an independent confirmation of mRNA copy number per cell, we compared the expression level of *SUP44/RPS4*, one of the few genes whose absolute mRNA levels have been reliably determined by quantitative hybridization experiments (Iyer and Struhl, 1996), with expression levels determined by SAGE. *SUP44/RPS4* was measured by hybridization at 75 +/- 10 copies/cell (Iyer and Struhl, 1996), in good accord with the SAGE data of 63 copies/cell, suggesting that the estimate of 15,000 mRNA molecules per cell was reasonably accurate. Analysis of SAGE tags from S phase arrested and G2/M phase arrested cells revealed similar expression levels for this gene (range 52 to 55 copies/cell), as well as for the vast majority of expressed genes. As less than 1% of the genes were expressed at dramatically different levels among these three states (see below), SAGE tags obtained from all libraries were combined and used to analyze global patterns of gene expression.

Analysis of ascertained tags at increasing increments revealed that the number of unique transcripts plateaued at ~60,000 tags (Figure 2). This suggested that generation of further SAGE tags would yield few additional genes, consistent with the fact that sixty thousand transcripts represented a four-fold redundancy for genes expressed as low as 1 transcript per cell.

Likewise, Monte Carlo simulations indicated that analysis of 60,000 tags would identify at least one tag for a given transcript 97% of the time if its expression level was one copy per cell.

The 56,291 tags that precisely matched the yeast genome represented 4,665 different genes. This number is in agreement with the estimate of 3,000 to 4,000 expressed genes obtained by RNA-DNA reassociation kinetics (Hereford and Rosbash, 1977). These expressed genes included 85% of the genes with characterized functions (1,981 of 2,340), and 76% of the total genes predicted from analysis of the yeast genome (4,665 of 6,121). These numbers are consistent with a relatively complete sampling of the yeast transcriptome given the limited number of physiological states examined and the large number of genes predicted solely on the basis of genomic sequence analysis.

The transcript expression per gene was observed to vary from 0.3 to over 200 copies per cell. Analysis of the distribution of gene expression levels revealed several abundance classes that were similar to those observed in previous studies using reassociation kinetics. A "virtual Rot" of the genes observed by SAGE (Figure 3A) identified three main components of the transcriptome with abundances ranging over three orders of magnitude. A Rot curve derived from RNA-cDNA reassociation kinetics also contained three main components distributed over a similar range of abundances (Hereford and Rosbash, 1977). Although the kinetics of reassociation of a particular class of RNA and cDNA may be affected by numerous experimental variables, there were striking similarities between Rot and virtual Rot analyses (Figure 3B). Because Rot analysis may not detect all transcripts of low abundance (Lewin, 1980), it is not surprising that SAGE revealed both a larger total number of expressed genes and a higher fraction of the transcriptome belonging to the low abundance transcript class.

Integration of Expression Information with the Genomic Map

The SAGE expression data could be integrated with existing positional

information to generate chromosomal expression maps (Figure 4). These maps were generated using the sequence of the yeast genome and the position coordinates of ORFs obtained from the Stanford Yeast Genome Database. Although there were a few genes that were noted to be physically proximal and have similarly high levels of expression, there did not appear to be any clusters of particularly high or low expression on any chromosome. Genes like histones H3 and H4, which are known to have coregulated divergent promoters and are immediately adjacent on chromosome 14 (Smith and Murray, 1983), had very similar expression levels (5 and 6 copies per cell, respectively). The distribution of transcripts among the chromosomes suggested that overall transcription was evenly dispersed, with total transcript levels being roughly linearly related to chromosome size (r^2 =0.85, data not shown). However, regions within 10 kb of telomeres appeared to be uniformly undertranscribed, containing on average 3.2 tags per gene as compared with 12.4 tags per gene for non-telomeric regions (Figure 4). This is consistent with the previously described observations of "telomeric silencing" in yeast (Gottschling et al., 1990). Recent studies have reported telomeric position effects as far as 4 kb from telomere ends (Renauld et al., 1993).

Gene Expression Patterns

Table 1 lists the 30 most highly expressed genes, all of which are expressed at greater than 60 mRNA copies per cell. As expected, these genes mostly correspond to well characterized enzymes involved in energy metabolism and protein synthesis and were expressed at similar levels in all three growth states (Examples in Figure 5). Some of these genes, including *ENO2* (McAlister and Holland, 1982), *PDC1* (Schmitt et al., 1983), *PGK1* (Chambers et al., 1989), *PYK1* (Nishizawa et al., 1989), and *ADH1* (Denis et al., 1983), are known to be dramatically induced in the glucose-rich growth conditions used in this study. In contrast, glucose repressible genes such as the *GAL1/GAL7/GAL10* cluster (St John and Davis, 1979), and *GAL3* (Bajwa

et al., 1988) were observed to be expressed at very low levels (0.3 or fewer copies per cell). As expected for the yeast strain used in this study, mating type a specific genes, such as the a factor genes (*MFA1, MFA2*) (Michaelis and Herskowitz, 1988), and alpha factor receptor (*STE2*) (Burkholder and Hartwell, 1985) were all observed to be expressed at significant levels (range 2 to 10 copies per cell), while mating type alpha specific genes (*MFα1, MFα2, STE3*) (Hagen et al., 1986; Kurjan and Herskowitz, 1982; Singh et al., 1983) were observed to be expressed at very low levels (<0.3 copies/cell).

Three of the highly expressed genes in Table 1 had not been previously characterized. One contained an ORF with predicted ribosomal function, previously identified only by genomic sequence analysis. Analyses of all SAGE data suggested that there were 2,684 such genes corresponding to uncharacterized ORFs which were transcribed at detectable levels. The 30 most abundant of these transcripts were observed more than 30 times, corresponding to at least 8 transcripts per cell (Table 2). The other two highly expressed uncharacterized genes corresponded to ORFs not predicted by analysis of the yeast genome sequence (NORF = Nonannotated ORF). Analyses of SAGE data suggested that there were approximately 160 *NORF* genes transcribed at detectable levels. The 30 most abundant of these transcripts were observed at least 9 times (Table 3 and examples in Figure 5).

Interestingly, one of the *NORF* genes (*NORF5*) was only expressed in S phase arrested cells and corresponded to the transcript whose abundance varied the most in the three states analyzed (> 49 fold, Figure 5). Comparison of S phase arrested cells to the other states also identified greater than 9 fold elevation of the *RNR2* and *RNR4* transcripts (Figure 5). Induction of these ribonucleoside reductase genes is likely to be due to the hydroxyurea treatment used to arrest cells in S phase (Elledge and Davis, 1989). Likewise, comparison of G2/M arrested cells identified elevation of RBL2 and dynein light chain, both microtubule associated proteins (Archer et al., 1995; Dick et al., 1996). As with the RNR inductions, these elevated levels seem likely to be related to the nocodazole treatment used to arrest cells in

the G2/M phase. While there were many relatively small differences between the states (for example, *NORF1*, Figure 5), overall comparison of the three states revealed surprisingly few dramatic differences; there were only 29 transcripts whose abundance varied more than 10 fold among the three different states analyzed.

Discussion

Analysis of a yeast transcriptome affords a unique view of the RNA components defining cellular life. We observed gene expression levels to vary over three orders of magnitude, with the transcripts involved in energy metabolism and protein synthesis the most highly expressed. Key transcripts, such as those encoding enzymes required for DNA replication (e.g. *POL1* and *POL3*), kinetochore proteins (*NDC10* and *SKP1*), and many other interesting proteins, were present at 1 or fewer copies per cell on average. These abundances are consistent with previous qualitative data from reassociation kinetics which suggested that the largest number of expressed genes was present at 1 or 2 copies per cell. These observations indicate that low transcript copy numbers are sufficient for gene expression in yeast, and suggest that yeast possess a mechanism for rigid control of RNA abundance.

The synthesis of chromosomal expression maps presents a cataloging of the expression level of genes, organized by their genomic positions. It is not surprising that gene expression is well balanced throughout the 16 chromosomes of *S. cerevisiae*. As most genes have independent regulatory elements, it would have been surprising to find a large number of physically adjacent genes that had similar high levels of expression. Of the few genes that were known to have coregulated divergent promoters, like the H3/H4 pair, SAGE data confirmed concordant levels of expression. For areas like telomere ends that are known to be transcriptionally suppressed, SAGE data corroborated low levels of expression. Other expected expression patterns such as high levels of glucose induced glycolytic enzymes, low levels of glucose repressed *GAL* genes, expression of mating type a specific genes, and

low of expression of mating type alpha genes, were observed. Finally, identification of tags corresponding to *NORF* genes suggests that there is a significant number of small proteins encoded by the yeast genome that were undetected by the criteria used for systematic sequence analysis. The yeast genome sequence has been annotated for all ORFS larger than 300bp, (encoding proteins 100 amino acids or greater). Genes encoding proteins below this cut off are therefore commonly unannotated. This class of genes might also be underrepresented in mutational collections because of the small target size for mutagenesis, and given their small size, may encode proteins with novel functions. The systematic knockout of these *NORF* genes will therefore be of great interest.

Comparison of gene expression patterns from altered physiologic states can provide insight into genes that are important in a variety of processes. Comparison of transcriptomes from a variety of physiologic states should provide a minimum set of genes whose expression is required for normal vegetative growth, and another set composed of genes that will be expressed only in response to specific environmental stimuli, or during specialized processes. For example, recent work has defined a minimal set of 250 genes required for prokaryotic cellular life (Mushegian and Koonin, 1996). Examination of the yeast genome readily identified homologous genes for 196 of these, over 90% of which were observed to be expressed in the SAGE analysis. Detailed analyses of yeast transcriptomes, as well as transcriptomes from other organisms, should ultimately allow the generation of a minimal set of genes required for eukaryotic life.

Like other genome-wide analyses, SAGE analysis of yeast transcriptomes has several potential limitations. First, a small number of transcripts would be expected to lack an NlaIII site and therefore would not be detected by our analysis. Second, our analysis was limited to transcripts found at least as frequently as 0.3 copies per cell. Transcripts expressed in only a minute fraction of the cell cycle, or transcripts expressed in only a fraction of the cell population, would not be reliably detected by our analysis.

Finally, mRNA sequence data are practically unavailable for yeast, and consequently, some SAGE tags cannot be unambiguously matched to corresponding genes. Tags which were derived from overlapping genes, or genes which have unusually long 3' untranslated regions may be misassigned. Increased availability of 3' UTR sequences in yeast mRNA molecules should help to resolve the ambiguities.

Despite these potential limitations, it is clear that the analyses described here furnish both global and local pictures of gene expression, precisely defined at the nucleotide level. These data, like the sequence of the yeast genome itself, provide simple, basic information integral to the interpretation of many experiments in the future. The availability of mRNA sequence information from EST sequencing as well as various genome projects, will soon allow definition of transcriptomes from a variety of organisms, including human. The data recorded here suggest that a reasonably complete picture of a human cell transcriptome will require only about 10 - 20 fold more tags than evaluated here, a number well within the practical realm achievable with a small number of automated sequencers. The analysis of global expression patterns in higher eukaryotes is expected, in general, to be similar to those reported here for *S. cerevisiae*. However, the analysis of the transcriptome in different cells and from different individuals should yield a wealth of information regarding gene function in normal, developmental, and disease states.

Experimental Procedures

Yeast cell culture

The source of transcripts for all experiments was S. cerevisiae strain YPH499 (*MATa ura3-52 lys2-801 ade2-101 leu2-Δ1 his3-Δ200 trp1-Δ63*) (Sikorski and Hieter, 1989). Logarithmically growing cells were obtained by growing yeast cells to early log phase (3 x 10^6 cells/ml) in YPD (Rose et al., 1990) rich medium (YPD supplemented with 6mM uracil, 4.8 mM adenine and 24 mM tryptophan) at 30°C. For arrest in the G1/S phase of the cell cycle,

hydroxyurea (0.1M) was added to early log phase cells, and the culture was incubated an additional 3.5 hours at 30°C. For arrest in the G2/M phase of the cell cycle, nocodazole (15ug/ml) was added to early log phase cells and the culture was incubated for an additional 100 minutes at 30°C. Harvested cells were washed once with water prior to freezing at -70°C. The growth states of the harvested cells were confirmed by microscopic and flow cytometric analyses (Basrai et al., 1996).

RNA isolation and Northern Blot Analysis

Total yeast RNA was prepared using the hot phenol method as described (Leeds et al., 1991). mRNA was obtained using the MessageMaker Kit (Gibco/BRL) following the manufacturer's protocol. Northern blot analysis was performed as described (El-Deiry et al., 1993), using probes PCR amplified from yeast genomic DNA.

SAGE protocol

The SAGE method was performed as previously described (Velculescu et al., 1995), with exceptions noted below. PolyA RNA was converted to double-stranded cDNA with a BRL synthesis kit using the manufacturer's protocol except for the inclusion of primer biotin-5'-T_{18}-3'. The cDNA was cleaved with NlaIII (Anchoring Enzyme). As NlaIII sites were observed to occur once every 309 base pairs in three arbitrarily chosen yeast chromosomes (1, 5, 10), 95% of yeast transcripts were predicted to be detectable with a NlaIII-based SAGE approach. After capture of the 3' cDNA fragments on streptavidin coated magnetic beads (Dynal), the bound cDNA was divided into two pools, and one of the following linkers containing recognition sites for BsmFI was ligated to each pool: Linker 1, 5'-TTTGGATTTGCTGGTGCAGTACAACTAGGCTTAATAGGGACATG-3' (SED ID NO:1). 5'-TCCCTATTAAGCCTAGTTGTACTGCACCAGCAAATCC [amino mod. C7]-3'(SED ID NO:2).; Linker 2,5'-

TTTCTGCTCGAATTCAAGCTTCTAACGATGTACGGGGACATG-3' (SED ID NO:3) 5'-TCCCCGTACATCGTTAGAAGCTTGAATTCGAGCAG[amino mod. C7]-3' (SED ID NO:4).

As BsmFI (Tagging Enzyme) cleaves 14 bp away from its recognition site, and the NlaIII site overlaps the BsmFI site by 1 bp, a 15 bp SAGE tag was released with BsmFI. SAGE tag overhangs were filled-in with Klenow, and tags from the two pools were combined and ligated to each other. The ligation product was diluted and then amplified with PCR for 28 cycles with 5'-GGATTTGCTGGTGCAGTACA-3' (SED ID NO:5) and 5'-CTGCTCGAATTCAAGCTTCT-3' (SED ID NO:6), as primers. The PCR product was analyzed by polyacrylamide gel electrophoresis (PAGE), and the PCR product containing two tags ligated tail to tail (ditag) was excised. The PCR product was then cleaved with NlaIII, and the band containing the ditags was excised and self-ligated. After ligation, the concatenated products were separated by PAGE and products between 500 bp and 2 kb were excised. These products were cloned into the SphI site of pZero (Invitrogen). Colonies were screened for inserts by PCR with M13 forward and M13 reverse sequences located outside the cloning site as primers.

PCR products from selected clones were sequenced with the TaqFS DyePrimer kits (Perkin Elmer) and analyzed using a 377 ABI automated sequencer (Perkin Elmer), following the manufacturer's protocol. Each successful sequencing reaction identified an average of 26 tags; given a 90% sequencing reaction success rate, this corresponded to an average of about 850 tags per sequencing gel.

SAGE data analysis

Sequence files were analyzed by means of the SAGE program group (Velculescu et al., 1995), which identifies the anchoring enzyme site with the proper spacing and extracts the two intervening tags and records them in a database. The 68,691 tags obtained contained 62,965 tags from unique

ditags and 5,726 tags from repeated ditags. The latter were counted only once to eliminate potential PCR bias of the quantitation, as described (Velculescu et al., 1995). Of 62,965 tags, 2,332 tags corresponded to linker sequences, and were excluded from further analysis. Of the remaining tags, 4,342 tags could not be assigned, and were likely due to sequencing errors (in the tags or in the yeast genomic sequence). If all of these were due to tag sequencing errors, this corresponds to a sequencing error rate of about 0.7% per base pair (for a 10bp tag), not far from what we would have expected under our automated sequencing conditions. However, some unassigned tags had a much higher than expected frequency of A's as the last five base pairs of the tag (5 of the 52 most abundant unassigned tags), suggesting that these tags were derived from transcripts containing anchoring enzyme sites within several base pairs from their polyA tails. Given the frequency of NlaIII sites in the genome (one in 309 base pairs), approximately 3% of transcripts were predicted to contain NlaIII sites within 10 bp of their polyA tails.

As very sparse data are available for yeast mRNA sequences and efforts to date have not been able to identify a highly conserved polyadenylation signal (Irniger and Braus, 1994; Zaret and Sherman, 1982), we used 14 bp of SAGE tags (i.e. the NlaIII site plus the adjacent 10 bp) to search the yeast genome directly (yeast genome sequence obtained from the Stanford yeast genome ftp site (genome-ftp.stanford.edu) on August 7, 1996). Because only coding regions are annotated in the yeast genome, and SAGE tags can be derived from 3' untranslated regions of genes, a SAGE tag was considered to correspond to a particular gene if it matched the ORF or the region 500 bp 3' of the ORF (locus names, gene names and ORF chromosomal coordinates were obtained from Stanford yeast genome ftp site, and ORF descriptions were obtained from MIPS www site (http://www.mips.biochem. mpg.de/) on August 14, 1996). ORFs were considered genes with known functions if they were associated with a three letter gene name, while ORFs without such designations were considered uncharacterized.

As expected, SAGE tags matched transcribed portions of the genome

in a highly non-random fashion, with 88% matching ORFs or their adjacent 3' regions in the correct orientation (chi-squared P value $<10^{-30}$). In instances when more than one tag matched a particular ORF in the correct orientation, the abundance was calculated to be the sum of the matched tags (for Figure 2, Figure 3, and Figure 4). Tags that matched ORFs in the incorrect orientation were not used in abundance calculations. In instances when a tag matched more than one region of the genome (for example an ORF and non-ORF region) only the matched ORF was considered. In some cases the 15th base of the tag could also be used to resolve ambiguities. For Figure 4, only tags that matched the genome once were used.

For the identification of NORF genes, only tags were considered that matched portions of the genome that were further than 500 bp 3' of a previously identified ORF, and were observed at least two times in the SAGE libraries.

Table 1. Highly expressed genes

Tag	Gene	Locus	Copies/cell	Description
GGTGTTAACG	TDH2/TDH3	YJR009C/YGR192C	425	glyceraldehyde-3-phosphate dehydrogenase 2 & 3
AGACAAACTG	TEF1/TEF2	YPR080W/YBR118W	248	cytosolic elongation factor eEF-1 alpha-A chain
TACCACTCCT	ENO2	YHR174W	229	2-phosphoglycerate dehydratase
GGTTTCGGTT	RPLA1, A2, A3, 10E	YDL081C/YOL039W/YDL130W/YLR340W	207	acidic ribosomal protein a1 / P2.beta / L44prime / L10
TTGCCAGTCT	PDC1	YLR044C	207	pyruvate decarboxylase isozyme 1
GGTGAAAACG	ADH1, ADH2	YOL086C/YMR303C	182	alcohol dehydrogenase I / II
ATCGCCGCTC	GPM1	YKL152C	168	phosphoglycerate mutase
GGTGCTAAGA	FBA1	YKL060C	166	fructose-bisphosphate aldolase II
TTAGTTTCTA	RPL47A	YDL184C	143	ribosomal protein
TCTCTACTGG	PGK1	YCR012W	139	phosphoglycerate kinase
GGTTTTGGTT	RPLA4	YDR382W	138	acidic ribosomal protein L45
GGTCCAGCTT	SSM1A / SSM1B	YPL220W / YGL135W	128	ribosomal protein
AATCCAGTTG	RPL5A / RPL5B	YIL018W / YFR031AC	102	ribosomal protein
TTCGTTCACT		NORF1	94	nonannotated ORF
AACAGACCAG	RPL16A / RPL16B	YPR102C / YGR085C	83	ribosomal protein
CTGCTCTGGG	CUP1A / CUP1B	YHR053C / YHR055C	75	metallothionein
GCAATACTAC		YOR293W / YMR230W	73	ribosomal protein S10 / similarity to ribosomal protein S10
GCTCTCCCCC		NORF2	73	nonannotated ORF
AAAGACAGAG	RPS31A	YGR027C	72	ribosomal protein
TGTCGTGGTG	RPL2A / RPL2B	YBR031W / YDR012W	70	ribosomal protein
CCAAGGGTAT	RPS28A	YGR118W	69	ribosomal protein
TCTCCAGAAG	RPL35B	YDR500C	69	ribosomal protein
GTTTTTCTTT	PYK1	YAL038W	69	pyruvate kinase
ATCACTGGTG	RPL9A / RPL9B	YGL147C / YNL067W	68	ribosomal protein L9
ATGAAGGTTC	RPL27A	YHR010W	68	ribosomal protein L27
GTAGAGCCGG	RPS21	YOL040C	67	ribosomal protein
GGTACTGATG	RPL43A	YDL075W	67	ribosomal protein L31
CCAGATTTGT	NAB1A / NAB1B	YGR214W / YLR048W	67	40S ribosomal protein p40 homolog A
GTGCCGTCCA	URP1A	YBR191W	62	ribosomal protein L21
CAAAACCCAA	RPS18EB	YML026C	60	ribosomal protein S18

Table 2. Putative coding sequences

SAGE Tag	Locus	Copies/Cell	Description
TTGAACTACC	YKL056C	58	strong similarity to human IgE-dependent histamine-releasing factor (21K tumor protein)
TTCGGGTCAC	YDR276C	56	strong similarity to Hordeum vulgare blt101 protein
CCAGATATGA	YIL093C	41	hypothetical protein
TTTAAAATGG	YMR116C	38	similarity to N.crassa CPC2 protein
GGTGTCGTTG	YBR078W	34	strong similarity to sporulation specific Sps2p
TACTCTTCGC	YEL033W	33	hypothetical protein
TGTAATTAAA	YOR182C	26	homology to human ubiquitin-like protein/ribosomal protein S30
GGAGATCTTG	YCR013C	24	weak similarity to M.lepra B1496_F1_41 protein
TCAAGAAGTT	YER056AC	20	strong similarity to ribosomal protein L34
AAAAACTTTG	YIL051C	18	strong similarity to YER057c
AAGTTGAACA	YPR043W	17	ribosomal protein L37
GGGTGCGGGT	YDR032C	16	strong similarity to YCR004c and S.pombe obr1
TGACTCTTTG	YLR390W	14	hypothetical protein
GGTCAATGGC	YJR105W	11	hypothetical protein
TAAGAATTCT	YJL158C	11	member of the Pir1p/Hsp150p/Pir3p family
TCAATTATGT	YDR033W	11	strong similarity to putative heat shock protein YRO2
ACGGCCAAGA	YBR162C	10	similarity to YJL171p
TTGGGCTAGT	YJL171C	10	similarity to YBR162c
CCTTCCAGGT	YJR085C	10	hypothetical protein
CCTCTCTTGT	YOR310C	10	homology to SIK1 protein
CCCAAAACTT	YEL018W	9	weak similarity to Rad50p
AACAAGTACT	YGL037C	9	similarity to E.coli hypothetical 23K protein
AACAATAAAA	YER072W	8	similarity to YFL004w
CAAAAGACCG	YML056C	8	homology to human IMP dehydrogenase I
GGTTTTTGAT	YOR182C	8	homology to human ubiquitin-like protein/ribosomal protein S30
CAATCCATTT	YBR106W	8	hypothetical protein
TTTTGGGTCT	YMR318C	8	putative alcohol-dehydrogenase
AACTGTCCAT	YDR429C	8	similarity to nuclear RNA binding proteins
CCAAGGTTAA	YAR002AC	8	strong similarity to YGL002w
GGTTTTTGAA	YOR273C	8	putative resistance protein

Table 3. NORF genes

SAGE Tag	Locus	Copies/Cell	Chr	Tag Pos	ORF Size (bp)
TTCGTTCACT	NORF1	94	4	1489450	198
GCTCTCCCCC	NORF2	73	16	75633	243
TGTACGCATT	NORF3	16	15	301251	189
TTTTATTATC	NORF4	15	6	223182	177
CTTCTCTTTT	NORF5	12	13	158973	204
TTTCCTATAA	NORF6	11	13	511754	252
TCTAGTCGCC	NORF7	10	12	669659	192
ATCGTTTTAT	NORF8	8	15	877140	174
GGCCAATGGT	NORF9	8	4	1202289	267
ACCCTGTCAT	NORF10	7	2	418633	255
AAAAGATCAT	NORF11	7	4	1489453	87
CAGAAAATGG	NORF12	6	8	115655	279
TGACATTCTT	NORF13	6	16	883669	183
TAGACATCTA	NORF14	6	2	491117	141
TGCCCTGGCC	NORF15	5	5	166452	216
GGTTTTGGCG	NORF16	4	3	24169	291
CCATACAGGT	NORF17	4	12	673851	114
CCAAATCAAA	NORF18	3	4	229494	258
AAGCGGTACT	NORF19	3	9	47889	399
AACGCTTTTC	NORF20	3	2	351456	198
GAGGATAGAG	NORF21	3	2	356201	240
CAATGAACCG	NORF22	3	16	75541	243
TCTTTATATA	NORF23	3	1	73363	90
CGCCTCCAGT	NORF24	3	7	485774	108
TACGTAAGTT	NORF25	3	10	156139	81
GATTTAAACT	NORF26	3	15	254749	93
GCGCCTCCAA	NORF27	2	5	42622	222
CAATGGCCCA	NORF28	2	13	511751	78
TTGAGGAACG	NORF29	2	3	154681	264
GCTAAGAACC	NORF30	2	4	302607	204

TABLE 4

Additional NORFs

SAGE Tag	Chr	Tag Pos	Copies/cell
GGCGCAATTT	4	1108395	2
TAAGTGATGA	7	593382	2
TTGTTGAATT	10	608373	2
GAAGCAGTAA	3	155607	2
ACATATGTTA	4	916112	2
CCCTACACGG	6	223289	2
GTAATTGGAC	10	392099	2
ATCAGACAAA	14	687272	2
TTATGAAAGA	15	81263	2
ATTCGTTCTA	15	841970	2
AGCAGGAGTT	16	188350	2
TTCTATTAGG	2	418749	2
TGGATTTCAG	4	1224930	2
CAGATATAAT	5	52488	2
CTGTTTTGGG	11	374761	2
CATTTTTAGT	11	508212	2
TTGAAAAGAT	13	104160	2
TAAGCCCATC	13	251273	2
AGCGTCCTCA	15	832420	2
TTTAGTTAAT	2	477623	2
ATGGTAGCCA	3	56961	2
AATTAGACTA	3	162589	2
AGTGACTCTT	4	1490879	2
GGACTATAAG	5	251266	2
ACTTTTTCAG	10	159213	2
GTCATATAGT	13	158765	2
CAACAAAGTG	13	171166	2
GTGGGAAAGG	13	804600	2
TACTTTATAT	16	366449	2
AATACCAGCG	3	175540	1
GCCTTGTATA	4	372624	1
GGTACATTCA	5	67152	1
GATTTCTCTG	5	187462	1
TAGTTGCTCC	7	317108	1
GTAAGAAATC	7	836202	1
CTTGGGCTAT	8	107992	1
AAATGGTGAT	11	558686	1
ATCATTTGGG	12	199358	1
CTGAACTTTA	12	283720	1
CCAGAAGGAG	13	652873	1
CCGGTTACTA	15	803663	1
CGATGAGAAG	15	1004369	1
AAACCGTCCC	16	199141	1
TCATTCATAC	2	164728	1
TATCTTTTTG	4	169784	1
TTAGAATAAT	4	603508	1
GTACGCTGTG	5	118089	1
TATATTAATT	6	64228	1

GTTCTTGCCT	7	939579	1
ATATAGCTGC	10	181144	1
CCAAAAAAAA	11	91785	1
GAACTCCACA	11	94125	1
CCTTCACTGC	11	374172	1
CACATCATAA	11	625896	1
GAAGTATTGA	12	603999	1
TGCGCGTATA	13	206410	1
GGGTAGTACT	13	671730	1
TAGTTTTGTC	15	33475	1
CAATTCCTAC	1	172182	0.8
TTTGATTTGA	2	46431	0.8
GGCTCTGGTT	2	414510	0.8
CAGAAATAGC	2	565130	0.8
CTGTTATTTT	2	616054	0.8
CGAAGTCAAA	2	680605	0.8
CTCTAGATAA	3	171584	0.8
AGTCAAAATG	4	192750	0.8
GCGAGTTTAG	4	691301	0.8
GCTCCAATAG	4	1131020	0.8
TTTATTTGAG	4	1237501	0.8
GTTATATTGA	4	1401803	0.8
TGGGTTGAAG	5	251266	0.8
ATTTTATTTG	5	447729	0.8
ATCATAAAAA	5	548612	0.8
TTATATAAAA	6	223182	0.8
CTACTTCTGC	8	34653	0.8
ATAAGACAGT	10	227802	0.8
TTCATAAGTT	10	471894	0.8
TAAATCTGAG	11	145617	0.8
CTGGTAGAAA	11	151174	0.8
CACGTACACA	11	403208	0.8
CCAAGATCAA	11	425882	0.8
AGCTTGTTCC	12	234966	0.8
CACATTCGTT	12	759953	0.8
CTTACATATA	12	789781	0.8
TCTATAGCAA	13	228936	0.8
CCTTTCTGAA	13	297985	0.8
CCTTTAGAAT	13	777999	0.8
AATTAACACC	13	842122	0.8
GCGCAGGGGC	14	440984	0.8
TGTTTATAAA	14	661710	0.8
AAAAGTCATT	15	32081	0.8
TTCGTAAACT	15	680625	0.8
TTTTTGGAGT	15	888343	0.8
AGGCATCTTG	16	250284	0.8
AAATCAAAAC	16	453890	0.8
AATTGACGAA	16	560169	0.8
TTGATGATTT	16	582360	0.8
CCTGTTTTTG	16	643476	0.8
TTTTTAAAAA	1	101436	0.5

AAGTTTGATC	1	199848	0.5
AGCACCTATG	2	46913	0.5
TGATTTATCC	2	418946	0.5
ACTGCATCTG	2	680860	0.5
CAAGTTAGGA	2	744770	0.5
ATACCCAATT	3	29939	0.5
AACTTTGTAT	3	30056	0.5
GCGGCGGGTG	3	41645	0.5
AAAATTGTTC	3	57108	0.5
TCAAGTACTC	3	157855	0.5
AACTGTATGC	3	223882	0.5
CTATCGGCCA	3	278840	0.5
ACAAGCCCAA	3	289917	0.5
GTACAGGGCT	4	93873	0.5
AAGATCATCG	4	254851	0.5
GAACTCCTGG	4	340891	0.5
GAACGAGAAG	4	371850	0.5
TTTTTAATAC	4	372058	0.5
TCTCCAGTTG	4	381712	0.5
AATACGTTAC	4	471791	0.5
ACGATTGGCT	4	509158	0.5
TGTTTATAAG	4	521709	0.5
CGTTTTCGTC	4	538839	0.5
TCGAACCTCT	4	578702	0.5
TCCACACACA	4	930972	0.5
CCGTGCGTGC	4	1324367	0.5
TTTCTTCAAC	5	116099	0.5
CCAAGTCTCG	5	159320	0.5
AGAGCGAATT	5	207517	0.5
TGTAGATTAT	5	280465	0.5
AAAAGTAGTT	5	286387	0.5
ACTTGGTATG	5	422942	0.5
TTAATGTTAT	5	544523	0.5
TACACGCGCG	5	544555	0.5
GGTCACTCCT	6	62983	0.5
AAGTGATGAA	6	76141	0.5
TTTATCTTGT	6	130327	0.5
AGTGATTGTT	6	256223	0.5
GCTTTGTTGT	7	72577	0.5
TCATTGATTC	7	110590	0.5
TTCACCGGAA	7	323655	0.5
ACTATTCTGT	7	423957	0.5
GGGCCAACCC	7	433787	0.5
AAAATATCTT	7	559397	0.5
TAGTAGTAAC	7	622201	0.5
AAGCGCACAA	7	735909	0.5
TCGCTGTTTT	7	800300	0.5
TGTATTTTTG	7	836202	0.5
CTAAACAAAG	7	836587	0.5
TAGGAAGAAA	7	905046	0.5
GGAAAAATTA	7	958839	0.5

TTTGGATAGT	7	974754	0.5
CGTTTGTGTA	8	202655	0.5
AGAAAAAAAC	8	386651	0.5
TAAAGTCCAG	8	518998	0.5
TAAGCAGATT	8	529129	0.5
ATGAGCATTT	9	97114	0.5
AGGTGCAAAA	9	229077	0.5
TAACAAAGAG	10	628227	0.5
CAATTGGCAA	10	721781	0.5
ACTCCCTGTA	11	93528	0.5
CTCTATTGAT	11	144281	0.5
GCTTTCCTTT	11	146665	0.5
ACCGCAAAGA	11	231872	0.5
CTTGTTCAAA	12	230972	0.5
AATGTGCTGT	12	320426	0.5
GCAGATAGCG	12	341324	0.5
TCTGACTTAG	12	368780	0.5
CCCGGATGTT	12	433912	0.5
GTAACGATTG	12	449917	0.5
GAATAACGAA	12	673851	0.5
ACTGCTATTT	12	712476	0.5
GTTCTCTAGC	12	712712	0.5
CATCACCATC	12	794710	0.5
TTGCACTTCT	12	806833	0.5
ACTGTTTATG	12	867350	0.5
TTGCTATATA	12	1017911	0.5
TACATTCTAA	13	95707	0.5
CTCTTAGTTG	13	158970	0.5
ACGAACACTT	13	278341	0.5
TGCGCAAGTC	13	283795	0.5
TTTTTCTTAA	13	363037	0.5
CAAATGCATT	13	390802	0.5
CAAATTGTGT	13	395599	0.5
GCAATACTAT	13	826521	0.5
AGTGACGATG	14	60143	0.5
TACTGGTTTA	14	118854	0.5
GTTTGACCTA	14	335512	0.5
AGCGTTTGAT	14	478481	0.5
CTCTGTTGCG	14	728251	0.5
AAATTCAAAA	15	35952	0.5
TTTGCTTGGT	15	242742	0.5
AGTTTTCCTG	15	304813	0.5
TTTAAAGATA	15	331453	0.5
AAGGAGACAC	15	448624	0.5
CTATATATCA	15	544530	0.5
GATGGAATAG	15	571210	0.5
TCGAGTCGAA	15	758202	0.5
AAAAAAGAAA	15	882567	0.5
TTTCCAGAAT	15	969884	0.5
TGGACAATGT	15	970607	0.5
GGAATTAAGA	15	979894	0.5

ACTATATGTT	16	582230	0.5
GATATATCAT	16	589647	0.5
AGAATTGATT	16	744406	0.5
CACTGTCTCC	16	824649	0.5

References

Archer, J. E., Vega, L. R., and Solomon, F. (1995). Rbl2p, a yeast protein that binds to beta-tubulin and participates in microtubule function in vivo. Cell *82*, 425-434.

Bajwa, W., Torchia, T. E., and Hopper, J. E. (1988). Yeast regulatory gene GAL3: carbon regulation; UASGal elements in common with GAL1, GAL2, GAL7, GAL10, GAL80, and MEL1; encoded protein strikingly similar to yeast and Escherichia coli galactokinases. Mol Cell Biol *8*, 3439-3447.

Basrai, M. A., Kingsbury, J., Koshland, D., Spencer, F., and Hieter, P. (1996). Faithful chromosome transmission requires Spt4p, a putative regulator of chromatin structure in Saccharomyces cerevisiae. Mol Cell Biol *16*, 2838-2847.

Bishop, J. O., Morton, J. G., Rosbash, M., and Richardson, M. (1974). Three abundance classes in HeLa cell messenger RNA. Nature *250*, 199-204.

Burkholder, A. C., and Hartwell, L. H. (1985). The yeast alpha-factor receptor: structural properties deduced from the sequence of the STE2 gene. Nucleic Acids Res *13*, 8463-8475.

Chambers, A., Tsang, J. S., Stanway, C., Kingsman, A. J., and Kingsman, S. M. (1989). Transcriptional control of the Saccharomyces cerevisiae PGK gene by RAP1. Mol Cell Biol *9*, 5516-5524.

Denis, C. L., Ferguson, J., and Young, E. T. (1983). mRNA levels for the fermentative alcohol dehydrogenase of Saccharomyces cerevisiae decrease upon growth on a nonfermentable carbon source. J Biol Chem *258*, 1165-1171.

Dick, T., Surana, U., and Chia, W. (1996). Molecular and genetic characterization of SLC1, a putative Saccharomyces cerevisiae homolog of the metazoan cytoplasmic dynein light chain1. Mol Gen Genet *251*, 38-43.

El-Deiry, W. S., Tokino, T., Velculescu, V. E., Levy, D. B., Parsons, R., Trent, J. M., Lin, D., Mercer, W. E., Kinzler, K. W., and Vogelstein, B. (1993). WAF1, a potential mediator of p53 tumor suppression. Cell *75*, 817-825.

Elledge, S. J., and Davis, R. W. (1989). DNA damage induction of ribonucleotide reductase. Mol Cell Biol *9*, 4932-4940.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S.G. (1996). Life with 6000 genes. Science *274*, 546-567.

Gottschling, D. E., Aparicio, O. M., Billington, B. L., and Zakian, V. A. (1990). Position effect at S. cerevisiae telomeres: reversible repression of Pol II transcription. Cell *63*, 751-762.

Hagen, D. C., McCaffrey, G., and Sprague, G. F., Jr. (1986). Evidence the yeast STE3 gene encodes a receptor for the peptide pheromone a factor: gene sequence and implications for the structure of the presumed receptor. Proc Natl Acad Sci U S A *83*, 1418-1422.

Hereford, L. M., and Rosbash, M. (1977). Number and distribution of polyadenylated RNA sequences in yeast. Cell *10*, 453-462.

Irniger, S., and Braus, G. H. (1994). Saturation mutagenesis of a polyadenylation signal reveals a hexanucleotide element essential for mRNA

3' end formation in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A *91*, 257-261.

Iyer, V., and Struhl, K. (1996). Absolute mRNA levels and transcriptional initiation rates in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A *93*, 5208-5212.

Kurjan, J., and Herskowitz, I. (1982). Structure of a yeast pheromone gene (MF alpha): a putative alpha-factor precursor contains four tandem copies of mature alpha-factor. Cell *30*, 933-943.

Leeds, P., Peltz, S. W., Jacobson, A., and Culbertson, M. R. (1991). The product of the yeast UPF1 gene is required for rapid turnover of mRNAs containing a premature translational termination codon. Genes Dev *5*, 230 3-2314.

Lewin, B. (1980). Gene Expression *2*, (New York, New York: John Wiley and Sons), pp. 694-727.

McAlister, L., and Holland, M. J. (1982). Targeted deletion of a yeast enolase structural gene. Identification and isolation of yeast enolase isozymes. J Biol Chem *257*, 7181-7188.

Michaelis, S., and Herskowitz, I. (1988). The a-factor pheromone of Saccharomyces cerevisiae is essential for mating. Mol Cell Biol *8*, 1309-1318.

Mushegian, A. R., and Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc. Natl. Acad. Sci. USA *93*, 10268-10273.

Nguyen, C., Rocha, D., Granjeaud, S., Baldit, M., Bernard, K., Naquet, P., and Jordan, B. R. (1995). Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones. Genomics *29*, 207-216.

Nishizawa, M., Araki, R., and Teranishi, Y. (1989). Identification of an upstream activating sequence and an upstream repressible sequence of the pyruvate kinase gene of the yeast Saccharomyces cerevisiae. Mol Cell Biol *9*, 442-451.

Renauld, H., Aparicio, O. M., Zierath, P. D., Billington, B. L., Chhablani, S. K., and Gottschling, D. E. (1993). Silent domains are assembled continuously from the telomere and are defined by promoter distance and strength, and by SIR3 dosage. Genes Dev *7*, 1133-1145.

Rose, M. D., Winston, F., and P. Hieter. (1990). Methods in Yeast Genetics. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press), pp. 177.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467-470.

Schmitt, H. D., Ciriacy, M., and Zimmermann, F. K. (1983). The synthesis of yeast pyruvate decarboxylase is regulated by large variations in the messenger RNA level. Mol Gen Genet *192*, 247-252.

Sikorski, R. S., and Hieter, P. (1989). A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. Genetics *122*, 19-27.

Singh, A., Chen, E. Y., Lugovoy, J. M., Chang, C. N., Hitzeman, R. A., and Seeburg, P. H. (1983). Saccharomyces cerevisiae contains two discrete genes coding for the alpha-factor pheromone. Nucleic Acids Res *11*, 4049-4063.

Smith, M. M., and Murray, K. (1983). Yeast H3 and H4 histone messenger RNAs are transcribed from two non-allelic gene sets. J Mol Biol *169*, 641-661.

St John, T. P., and Davis, R. W. (1979). Isolation of galactose-inducible DNA sequences from Saccharomyces cerevisiae by differential plaque filter hybridization. Cell *16*, 443-452.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. Science *270*, 484-487.

Zaret, K. S., and Sherman, F. (1982). DNA sequence required for efficient transcription termination in yeast. Cell *28*, 563-573.

CLAIMS

1. An isolated DNA molecule comprising a yeast gene which is involved in cell cycle progression selected from the group of NORF genes identified in Tables 3 and 4.

2. The isolated DNA molecule of claim 1 wherein expression of the NORF gene varies by at least 10% between any two phases of the cell cycle selected from the group consisting of: log phase, S phase, and G2/M.

3. The isolated DNA molecule of claim 1 wherein expression of the NORF gene varies by at least 25% between any two phases of the cell cycle selected from the group consisting of: log phase, S phase, and G2/M.

4. The isolated DNA molecule of claim 1 wherein expression of the NORF gene varies by at least 50% between any two phases of the cell cycle selected from the group consisting of: log phase, S phase, and G2/M.

5. The isolated DNA molecule of claim 1 wherein expression of the NORF gene varies by at least 100% between any two phases of the cell cycle selected from the group consisting of: log phase, S phase, and G2/M.

6. The isolated DNA molecule of claim 1 wherein expression of the NORF gene varies by a statistically significant difference (greater than 95% confidence level) between any two phases of the cell cycle selected from the group consisting of: log phase, S phase, and G2/M.

7. The isolated DNA molecule of claim 6 wherein the NORF is selected from the group consisting of NORF Nº 1, 2, 4, 5, 6, 17, 25, and 27.

8. The isolated DNA molecule of claim 1 wherein the NORF gene is not expressed in at least one phase of the cell cycle selected from the group consisting of: log phase, S phase, and G2/M.

9. The isolated DNA molecule of claim 1 which is genomic.

10. The isolated DNA molecule of claim 1 which is cDNA.

11. A method of using yeast genes to affect the cell cycle, comprising the step of:

administering to a cell an isolated DNA molecule comprising a yeast gene which is involved in cell cycle progression selected from the differentially expressed genes identified in Tables 1, 2, 3, and 4.

12. The method of claim 11 wherein the cell is a yeast cell.

13. The method of claim 11 wherein the cell is a fungal cell.

14. The method of claim 11 wherein the cell is a mammalian cell.

15. The method of claim 11 wherein the yeast gene is selected from the group consisting of NORF Nº 1, 2, 4, 5, 6, 17, 25, and 27.

16. The method of claim 11 wherein the yeast gene is selected from the group consisting of: TEF1/TEF2, EN02, ADH1, ADH2, PGK1, CUP1A/CUP1B, and PYK1.

17. The method of claim 11 wherein the yeast gene is selected from the group consisting of: YKL056C, YMR116C, YEL033W, YOR182C, YCR013C, and YJR085C.

18. A method for screening candidate antifungal drugs, comprising the steps of:

contacting a test substance with a yeast cell;

monitoring expression of a yeast gene which is involved in cell cycle progression selected from the group of yeast genes identified in Tables 1, 2, 3, and 4, wherein a test substance which modifies the expression of the yeast gene is a candidate antifungal drug.

19. The method of claim 18 wherein the yeast gene is selected from the group consisting of NORF Nº 1, 2, 4, 5, 6, 17, 25, and 27.

20. The method of claim 18 wherein the yeast gene is selected from the group consisting of: TEF1/TEF2, EN02, ADH1, ADH2, PGK1, CUP1A/CUP1B, and PYK1.

21. The method of claim 18 wherein the yeast gene is selected from

the group consisting of: YKL056C, YMR116C, YEL033W, YOR182C, YCR013C, and YJR085C.

22. A method for identifying human genes which are involved in cell cycle progression, comprising the steps of:

hybridizing a probe comprising at least 10 contiguous nucleotides of a yeast gene which is differentially expressed between at least two phases selected from the group consisting of log phase, S phase, and G2/M phase, wherein the yeast gene is identified in Table 1, 2, 3, or 4.

23. The method of claim 22 wherein the yeast gene is selected from the group consisting of NORF Nº 1, 2, 4, 5, 6, 17, 25, and 27.

24. The method of claim 22 wherein the yeast gene is selected from the group consisting of: TEF1/TEF2, EN02, ADH1, ADH2, PGK1, CUP1A/CUP1B, and PYK1.

25. The method of claim 22 wherein the yeast gene is selected from the group consisting of: YKL056C, YMR116C, YEL033W, YOR182C, YCR013C, and YJR085C.

26. A probe for ascertaining phase in the cell cycle of a cell, wherein the probe comprises at least 14 contiguous nucleotides of a NORF gene as identified in Table 3 or 4.

27. The probe of claim 26 wherein expression of the NORF gene varies by at least 10% between any two phases of the cell cycle selected from the group consisting of: log phase, S phase, and G2/M.

28. The probe of claim 26 wherein expression of the NORF gene varies by at least 25% between any two phases of the cell cycle selected from the group consisting of: log phase, S phase, and G2/M.

29. The probe of claim 26 wherein expression of the NORF gene varies by at least 50% between any two phases of the cell cycle selected from the group consisting of: log phase, S phase, and G2/M.

30. The probe of claim 26 wherein expression of the NORF gene varies by at least 100% between any two phases of the cell cycle selected from the group consisting of: log phase, S phase, and G2/M.

31. The probe of claim 26 wherein the NORF gene is not expressed in at least one phase of the cell cycle selected from the group consisting of: log phase, S phase, and G2/M.

32. The probe of claim 26 wherein expression of the NORF gene varies by a statistically significant difference (greater than 95% confidence level) between any two phases of the cell cycle selected from the group consisting of: log phase, S phase, and G2/M.

33. The probe of claim 32 wherein the gene is selected from the group consisting of NORF Nº 1, 2, 4, 5, 6, 17, 25, and 27.

34. The method of claim 18 wherein said step of monitoring expression is performed using nucleic acid molecules which are immobilized on a solid support.

35. The method of claim 34 wherein the nucleic acid molecules are in on array.

36. The method of claim 19 wherein a probe which comprises a portion of said yeast gene is in an array on a solid support.

37. An array of probes on a solid support wherein at least one probe comprises at least 14 contiguous nucleotides of a NORF gene as identified in Table 3 or 4.

38. The array of claim 37 wherein the NORF gene is selected from the group consisting of NORF No. 1 2, 4, 5, 6, 17, 25, and 27.

39. The array of claim 37 which comprises at least 100 probes of distinct sequence .

40. The array of claim 37 which comprises at least 500 probes of distinct sequence.

41. The array of claim 37 which comprises at least 1,000 probes of distinct sequence.

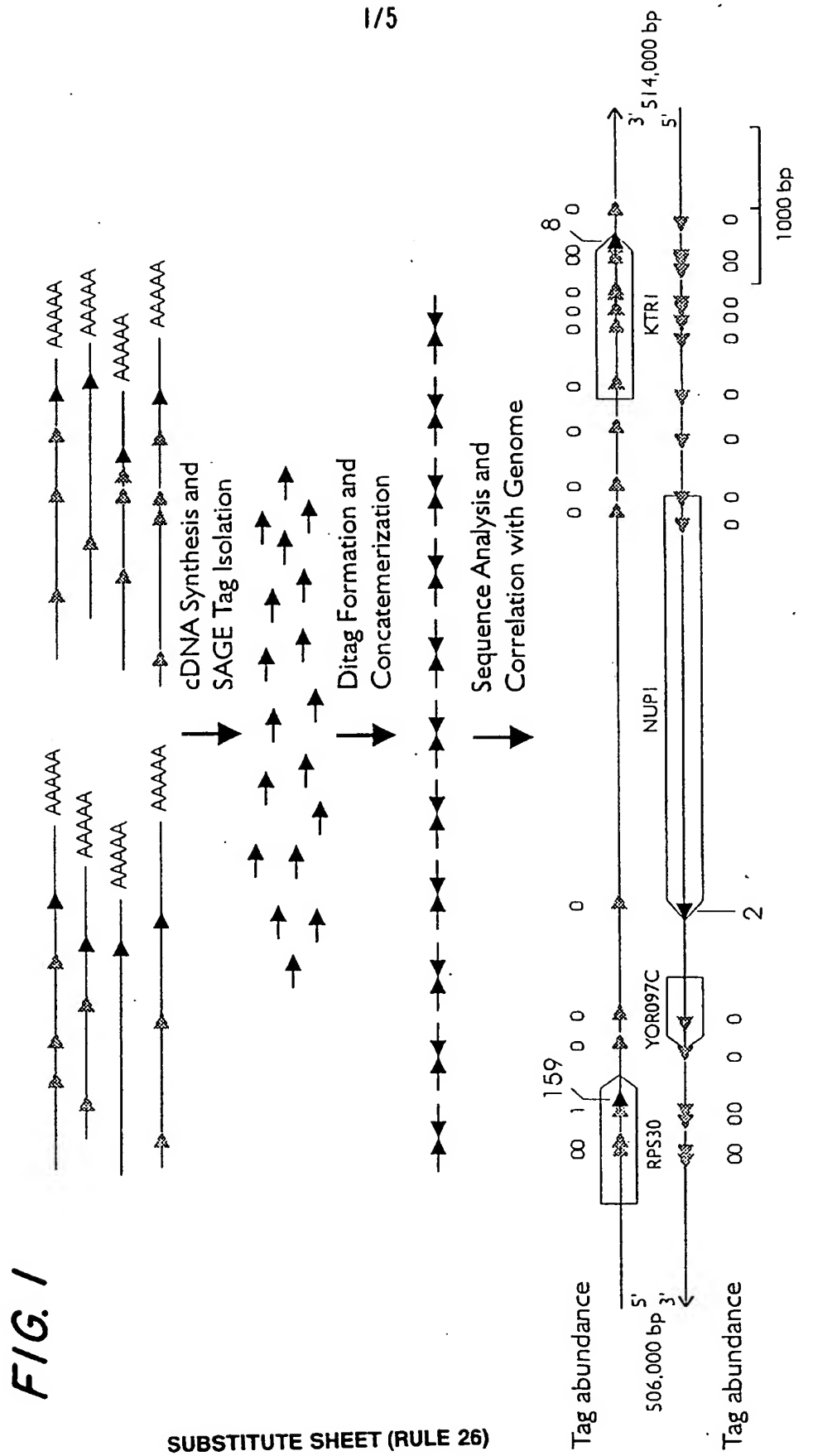

AAAAA
cDNA Synthesis and
SAGE Tag Isolation
Ditag Formation and
Concatemerization
Sequence Analysis and
Correlation with Genome
Tag abundance
506,000 bp
5'
3'
159
RPS30
YOR097C
NUP1
KTR1
514,000 bp
1000 bp
Tag abundance

FIG. 1

FIG. 2

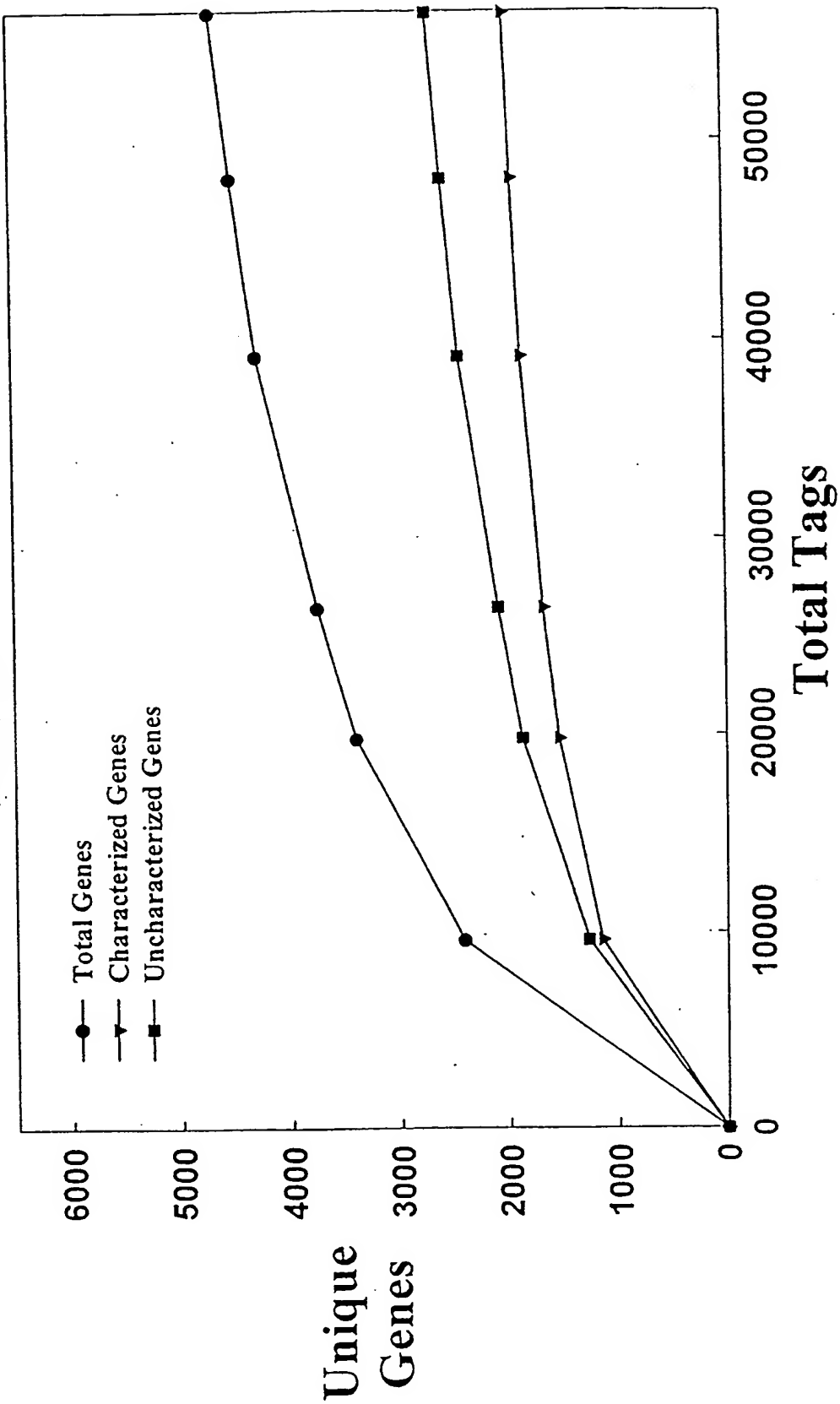

Total Genes
Characterized Genes
Uncharacterized Genes
Unique Genes
Total Tags
0
1000
2000
3000
4000
5000
6000
0
10000
20000
30000
40000
50000

FIG. 3A

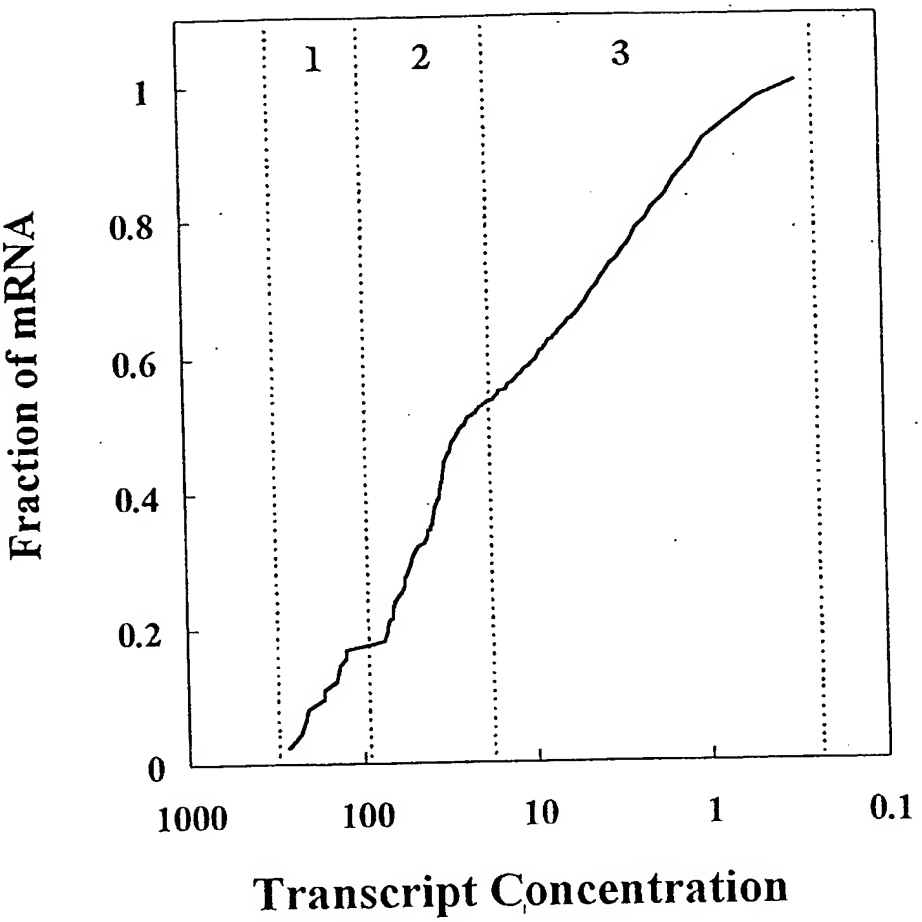


FIG. 3B

Component	Virtual Rot (SAGE) %mRNA	Virtual Rot (SAGE) Copies/cell	Rot (Reassociation) %mRNA	Rot (Reassociation) Copies/cell
1	17	180	23	200
2	38	40	51	30
3	45	2.5	26	1.5

FIG. 4

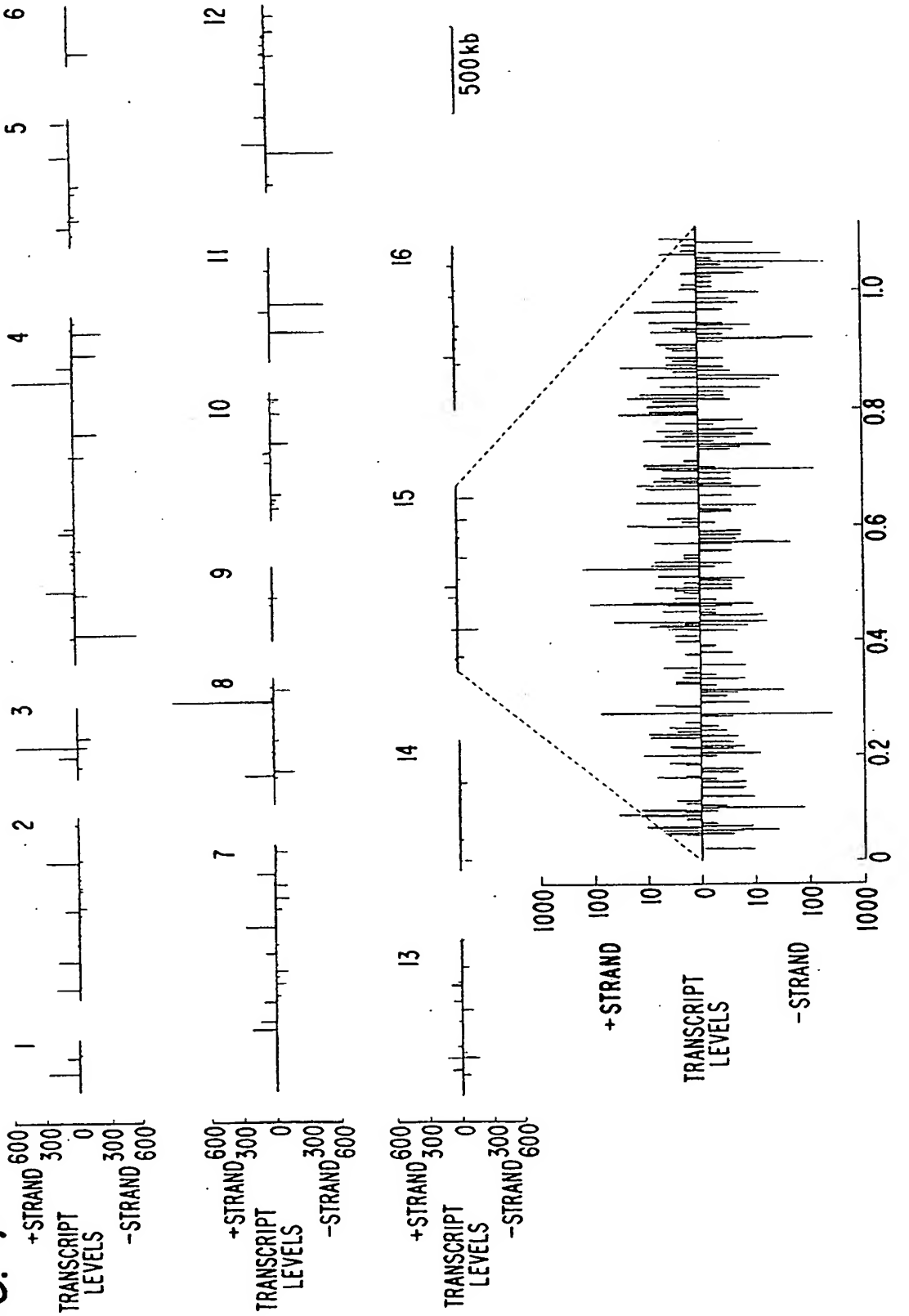
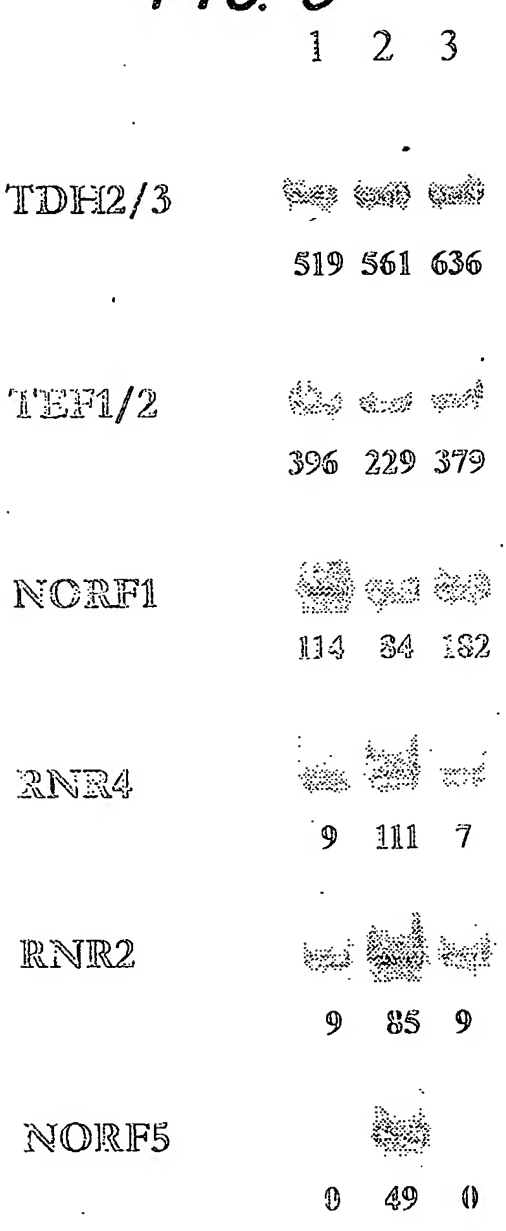

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
500kb
+STRAND
TRANSCRIPT LEVELS
-STRAND
600
300
0
300
600
1000
100
10
0
10
100
1000
0
0.2
0.4
0.6
0.8
1.0

FIG. 5


1 2 3
TDH2/3
519 561 636
TEF1/2
396 229 379
NORF1
114 84 182
RNR4
9 111 7
RNR2
9 85 9
NORF5
0 49 0